# Design Evaluation

[Preece Chap 29-34; Hix Chap 10; Neilsen Chap 6-7; Newman Chap 5,8-10]

## Why evaluate?

Understand the problem
Compare designs
Fine-tune an engineering solution
Checking performance against usability specifications

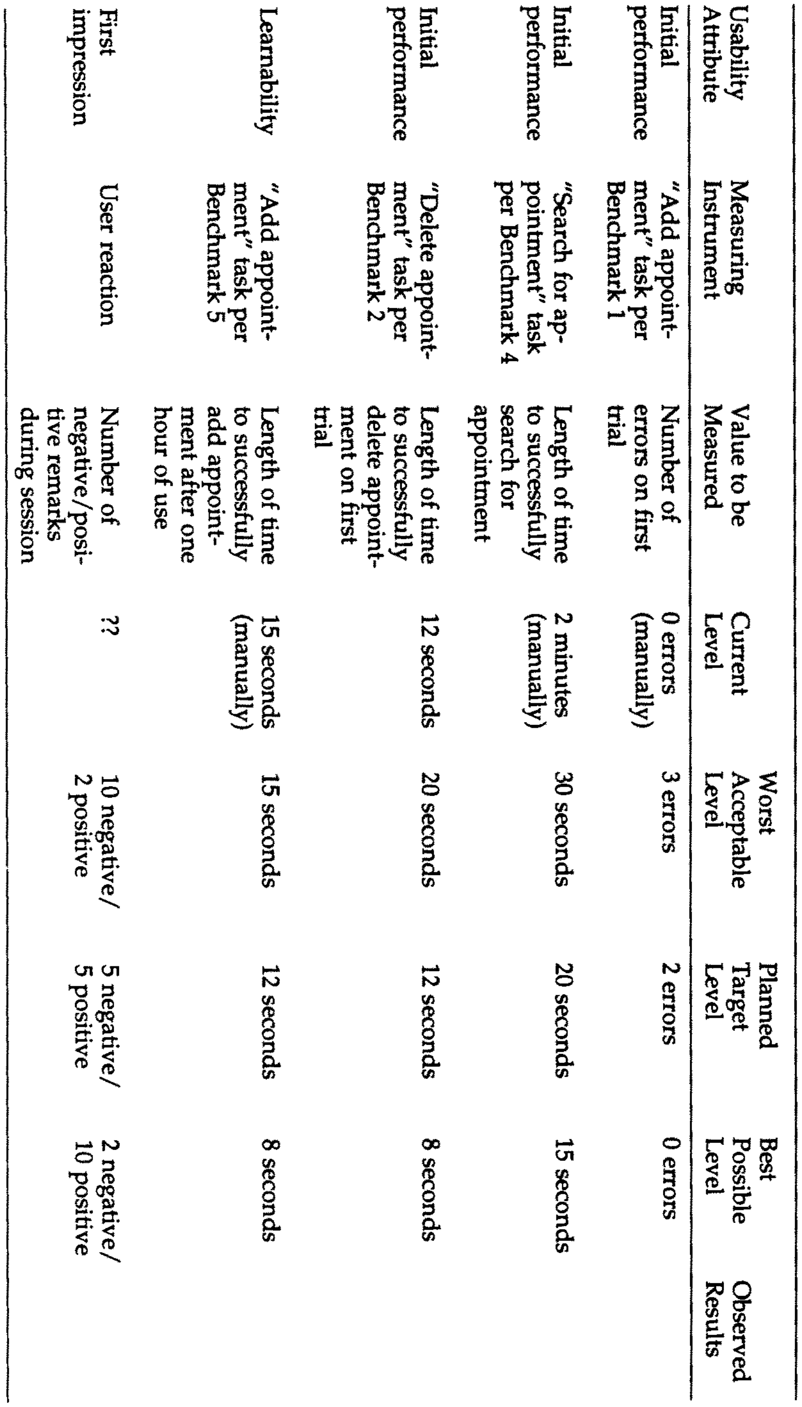
There are two fundamentally different types of evaluation we need to be concerned with.

**Informal evaluation.**
This is a typical usability study. It is most appropriate where the actual usage patterns of a system are somewhat unpredictable. If you wanted to compare two different interface designs to a word processor, there would be too great variability in usage between any two people so a formal evaluation would be of minimal interest. However, a usability study that compares performance to specification, error rates, etc., would be appropriate.

**Formal evaluation.**
In cases where usage is predictable, a formal evaluation may be suitable. This is used very often in science to determine fundamental principles for design. For example, if you wanted to know if a pulldown menu with ten items was better or worse than two pulldown menus with five items, a formal experiment would be used to find out. Formal evaluations follow typical scientific methods; they have a hypothesis, a formal design, etc. Informal evaluation do not necessarily follow these guidelines. Consequently, the results of a usability evaluation are usually only of interest to that design and do not transfer to others.

## Usability Specifications

When you perform an evaluation, how will you know if the level of usability is acceptable for general human use or not? You must determine beforehand what the usability specifications are. These are *measurable* so that the results of your evaluation can be compared against them.



| Usability Attribute | Measuring Instrument | Value to be Measured | Current Level | Worst Acceptable Level | Planned Target Level | Best Possible Level | Observed Results |
|---|---|---|---|---|---|---|---|
| Initial performance | "Add appointment" task per Benchmark 1 | Number of errors on first trial | 0 errors (manually) | 3 errors | 2 errors | 0 errors | |
| Initial performance | "Search for appointment" task per Benchmark 4 | Length of time to successfully search for appointment | 2 minutes (manually) | 30 seconds | 20 seconds | 15 seconds | |
| Initial performance | "Delete appointment" task per Benchmark 2 | Length of time to successfully delete appointment on first trial | 12 seconds | 20 seconds | 12 seconds | 8 seconds | |
| Learnability | "Add appointment" task per Benchmark 5 | Length of time to successfully add appointment after one hour of use | 15 seconds (manually) | 15 seconds | 12 seconds | 8 seconds | |
| First impression | User reaction | Number of negative/positive remarks during session | ?? | 10 negative/ 2 positive | 5 negative/ 5 positive | 2 negative/ 10 positive | |

Usability Attributes
include:

- Initial performance
- Long-term performance
- Learnability
- Retainability
- Advanced feature use
- First impression
- Long-term user satisfaction

The specification should provide a:

- worst acceptable level,
- a planned target level,
- a current level, and
- a best possible level for each item.

Reasonable measurable values include:

- Time to complete a task
- Number or percentage of errors
- Percentage of task completed in a given time
- Ratio of successes to failures
- Time spent in errors and recovery
- Number of commands/actions used to perform tasks
- Frequency of help and documentation use
- Number of repetitions of failed commands
- Number of available commands not invoked
- Number of times user expresses frustration or satisfaction

## Formative versus Summative Evaluation

*Formative evaluation* is a type of usability evaluation performed early and continuously throughout development; its purpose is to assess and improve the user interface design.

*Summative evaluation*, in contrast, is typically performed after a system or user interface is more or less complete; its purpose is to statistically compare several different systems or interfaces, for example, to determine which one is ‘better’ — where better is defined in advance.

Neither type of evaluation is more formal than the other; they just have different purposes. Formative evaluation is the type that ensures usability of interactive systems.

## Steps in Performing an Evaluation

1. Develop the experiment
2. Direct the evaluation session
3. Collect the data
4. Analyze the data
5. Draw conclusions to form a resolution for each identified design problem
6. Redesign and implement the revised interface

See Greenberg [Included]

## Develop the Experiment

- Select participants to perform the tasks
- Develop tasks for participants to perform
- Determine protocol and procedures for the evaluation sessions
- Pilot testing to shake down the experiment

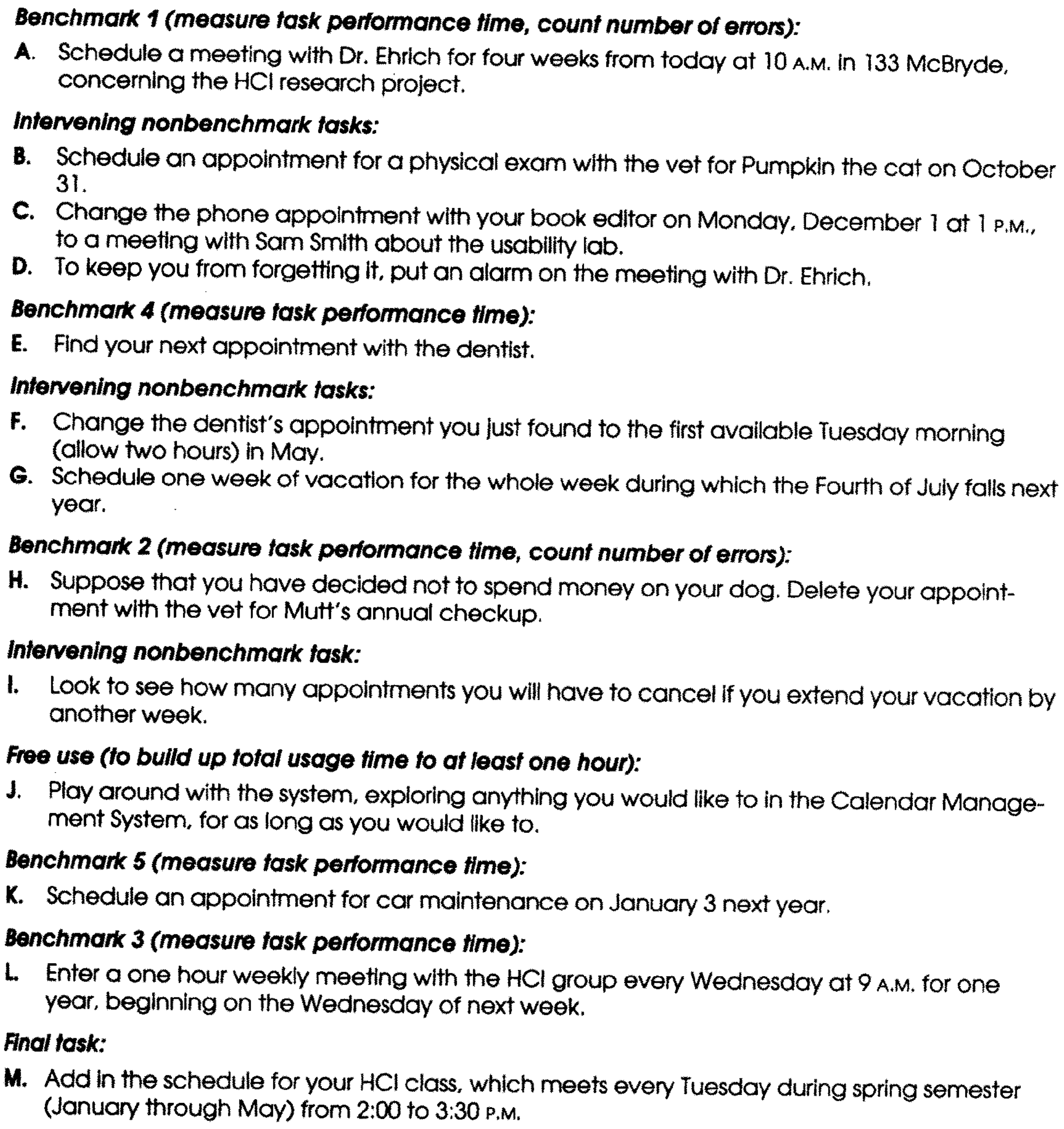
### Select Participants

Select a representative set of people who you think will be the typical users of the system. These people will give you your best feedback of what works and what doesn’t.

Be careful not to select participants that may know too much about the interface being studied. People who have an idea what you might be studying will behave differently from those who regard it as a simple working system. This does not imply that novice users are always best. It is often a good idea to select people who are at least a little familiar with the problem domain.

**Develop Tasks**

The evaluator's copy of the task list.



***Benchmark 1 (measure task performance time, count number of errors):***

**A.** Schedule a meeting with Dr. Ehrich for four weeks from today at 10 A.M. in 133 McBryde, concerning the HCI research project.

***Intervening nonbenchmark tasks:***

**B.** Schedule an appointment for a physical exam with the vet for Pumpkin the cat on October 31.

**C.** Change the phone appointment with your book editor on Monday, December 1 at 1 P.M., to a meeting with Sam Smith about the usability lab.

**D.** To keep you from forgetting it, put an alarm on the meeting with Dr. Ehrich.

***Benchmark 4 (measure task performance time):***

**E.** Find your next appointment with the dentist.

***Intervening nonbenchmark tasks:***

**F.** Change the dentist's appointment you just found to the first available Tuesday morning (allow two hours) in May.

**G.** Schedule one week of vacation for the whole week during which the Fourth of July falls next year.

***Benchmark 2 (measure task performance time, count number of errors):***

**H.** Suppose that you have decided not to spend money on your dog. Delete your appointment with the vet for Mutt's annual checkup.

***Intervening nonbenchmark task:***

**I.** Look to see how many appointments you will have to cancel if you extend your vacation by another week.

***Free use (to build up total usage time to at least one hour):***

**J.** Play around with the system, exploring anything you would like to in the Calendar Management System, for as long as you would like to.

***Benchmark 5 (measure task performance time):***

**K.** Schedule an appointment for car maintenance on January 3 next year.

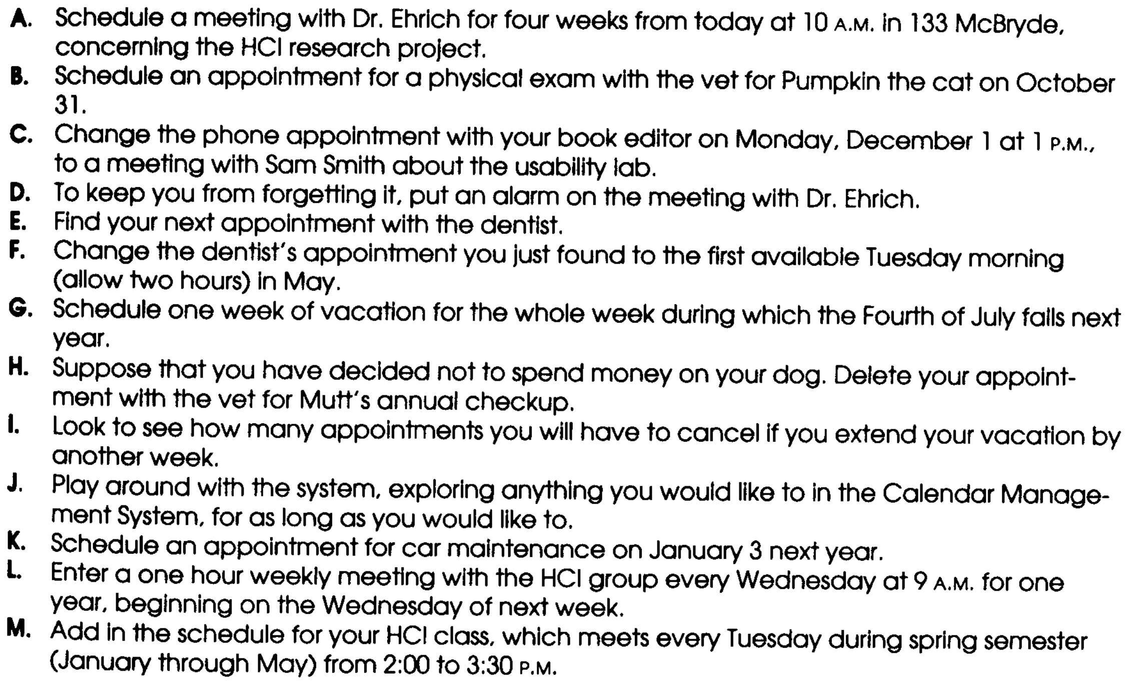
***Benchmark 3 (measure task performance time):***

**L.** Enter a one hour weekly meeting with the HCI group every Wednesday at 9 A.M. for one year, beginning on the Wednesday of next week.

***Final task:***

**M.** Add in the schedule for your HCI class, which meets every Tuesday during spring semester (January through May) from 2:00 to 3:30 P.M.

The participant's copy of the task list



**A.** Schedule a meeting with Dr. Ehrich for four weeks from today at 10 A.M. in 133 McBryde, concerning the HCI research project.
**B.** Schedule an appointment for a physical exam with the vet for Pumpkin the cat on October 31.
**C.** Change the phone appointment with your book editor on Monday, December 1 at 1 P.M., to a meeting with Sam Smith about the usability lab.
**D.** To keep you from forgetting it, put an alarm on the meeting with Dr. Ehrich.
**E.** Find your next appointment with the dentist.
**F.** Change the dentist's appointment you just found to the first available Tuesday morning (allow two hours) in May.
**G.** Schedule one week of vacation for the whole week during which the Fourth of July falls next year.
**H.** Suppose that you have decided not to spend money on your dog. Delete your appointment with the vet for Mutt's annual checkup.
**I.** Look to see how many appointments you will have to cancel if you extend your vacation by another week.
**J.** Play around with the system, exploring anything you would like to in the Calendar Management System, for as long as you would like to.
**K.** Schedule an appointment for car maintenance on January 3 next year.
**L.** Enter a one hour weekly meeting with the HCI group every Wednesday at 9 A.M. for one year, beginning on the Wednesday of next week.
**M.** Add in the schedule for your HCI class, which meets every Tuesday during spring semester (January through May) from 2:00 to 3:30 P.M.

**Determine Protocol and Procedures**
(See attached Human Subjects Procedure)

- Protocol: exactly what are you going to do
- Consent form: participants know what is expected of them
- Debriefing: background on the experiment

Within government supported research, these procedures must be adhered to when the use of human subjects is involved.

**Pilot Testing**
The earliest pilot testers are usually the design team or experimenters themselves. This is an acceptable practice to reach a reasonable level of operability and usability. You don't want to start with a system that is extremely far from an acceptable solution if it can be helped.
Subsequent pilot testing involves running one or two participants through the experimental procedure in an informal way to determine if the procedure is appropriate and if the data collected will satisfy the needs of the study.

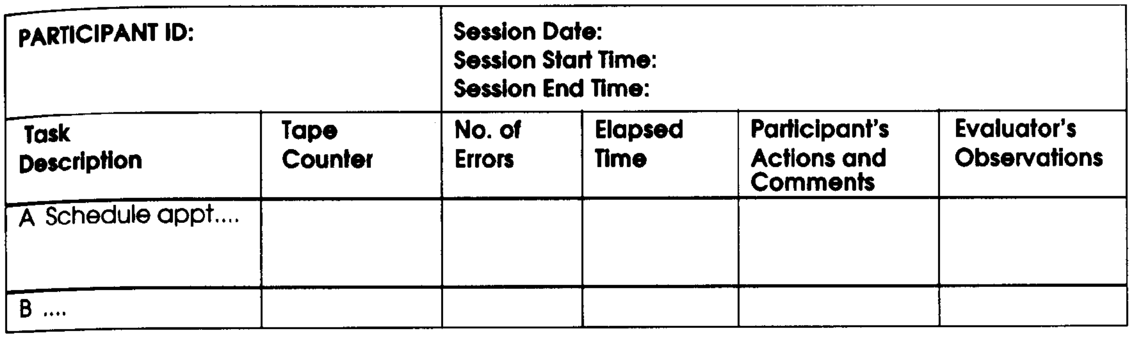
## Types of Evaluation Data

- *Objective*: These are directly observed measures, typically of user performance while using the interface to perform benchmark tasks.
- *Subjective*: These represent opinions, usually of the user, concerning usability of the interface.
- *Quantitative*: These are numeric data and results, such as user performance metrics or opinion ratings. This kind of data is key in helping to monitor convergence toward usability specifications during all cycles of iterative development.
- *Qualitative*: These are non-numeric data and results, such as lists of problems users had while using the interface, and they result in suggestions for modifications to improve the interaction design. This kind of data is useful in identifying which design features are associated with measured usability problems during all cycles of iterative development.

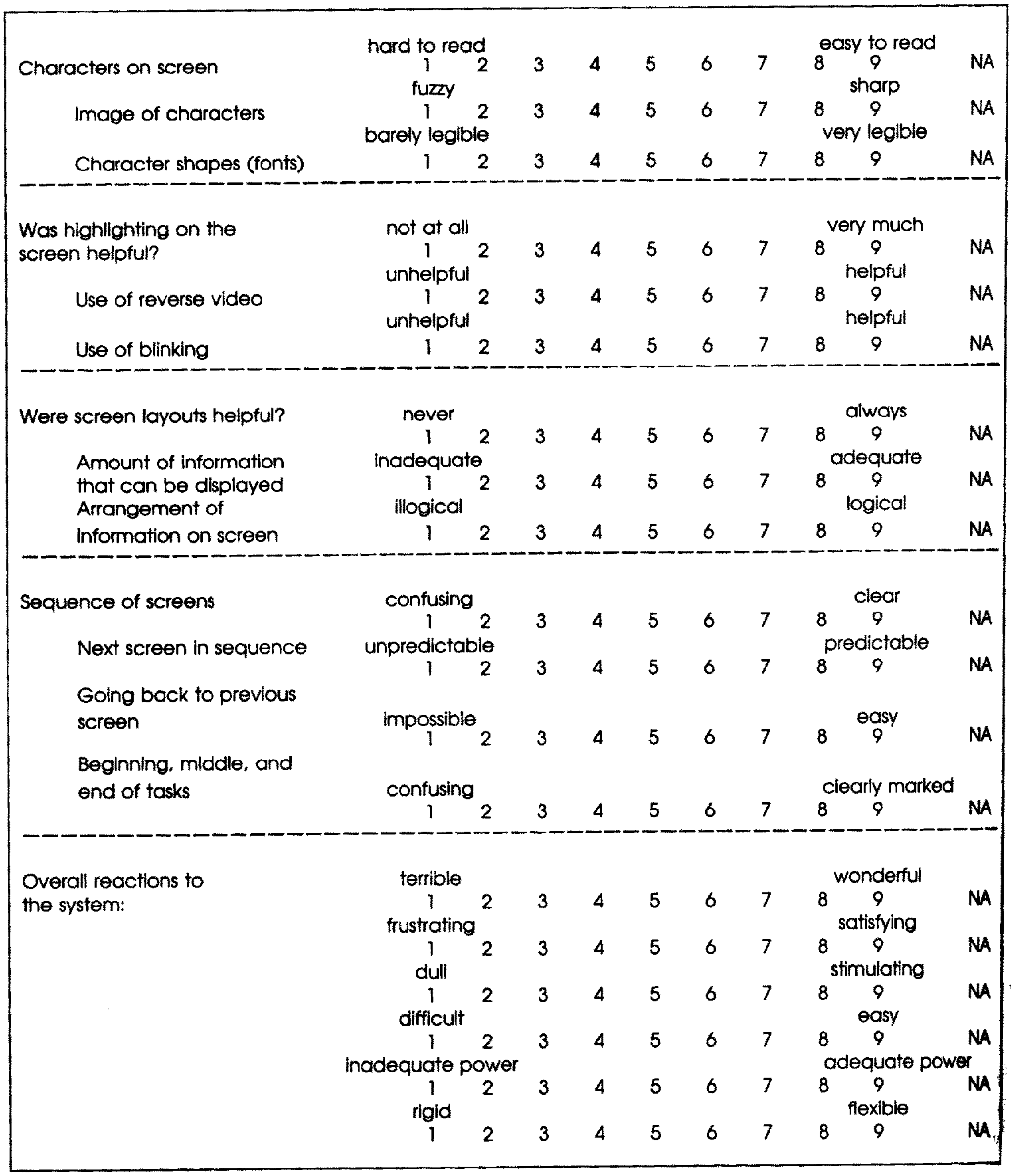
A common misconception is that quantitative measures are associated with objective measures and qualitative measures are associated with subjective measures. This is not the case.

## Quantitative Measures

- Benchmark tasks
- User questionnaires
- Post-Hoc Analyses (In-situ data logs)



| | | | | | |
|---|---|---|---|---|---|
| PARTICIPANT ID: | | Session Date:<br>Session Start Time:<br>Session End Time: | | | |
| Task Description | Tape Counter | No. of Errors | Elapsed Time | Participant's Actions and Comments | Evaluator's Observations |
| A Schedule appt.... | | | | | |
| B .... | | | | | |



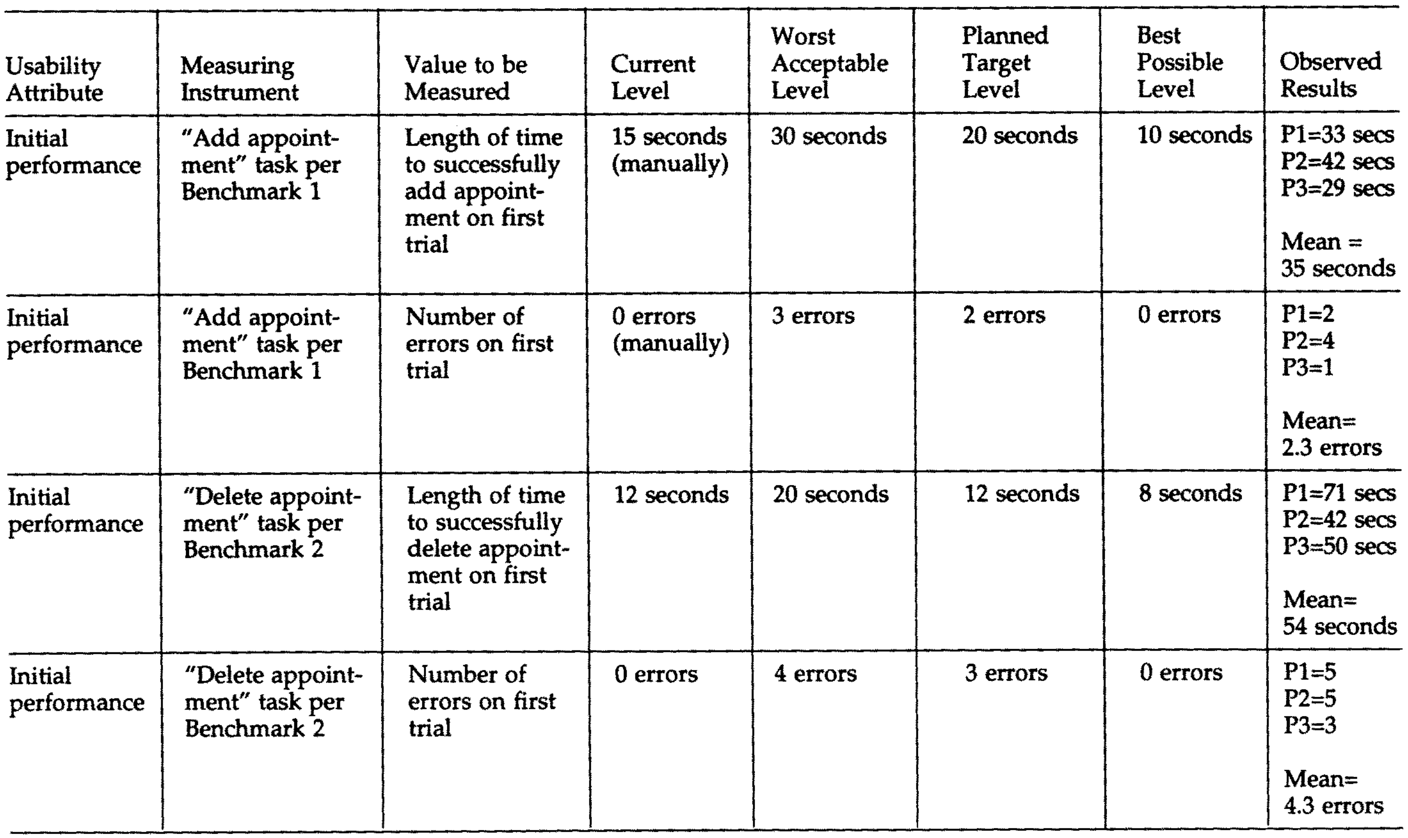
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Characters on screen | hard to read | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | easy to read | NA |
| Image of characters | fuzzy | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | sharp | NA |
| Character shapes (fonts) | barely legible | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | very legible | NA |
| Was highlighting on the screen helpful? | not at all | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | very much | NA |
| Use of reverse video | unhelpful | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | helpful | NA |
| Use of blinking | unhelpful | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | helpful | NA |
| Were screen layouts helpful? | never | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | always | NA |
| Amount of information that can be displayed | inadequate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | adequate | NA |
| Arrangement of information on screen | illogical | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | logical | NA |
| Sequence of screens | confusing | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | clear | NA |
| Next screen in sequence | unpredictable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | predictable | NA |
| Going back to previous screen | impossible | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | easy | NA |
| Beginning, middle, and end of tasks | confusing | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | clearly marked | NA |
| Overall reactions to the system: | terrible | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | wonderful | NA |
| | frustrating | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | satisfying | NA |
| | dull | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | stimulating | NA |
| | difficult | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | easy | NA |
| | inadequate power | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | adequate power | NA |
| | rigid | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | flexible | NA |

## Qualitative Measures

- Direct observation
- Concurrent verbal protocols
- Retrospective verbal protocols

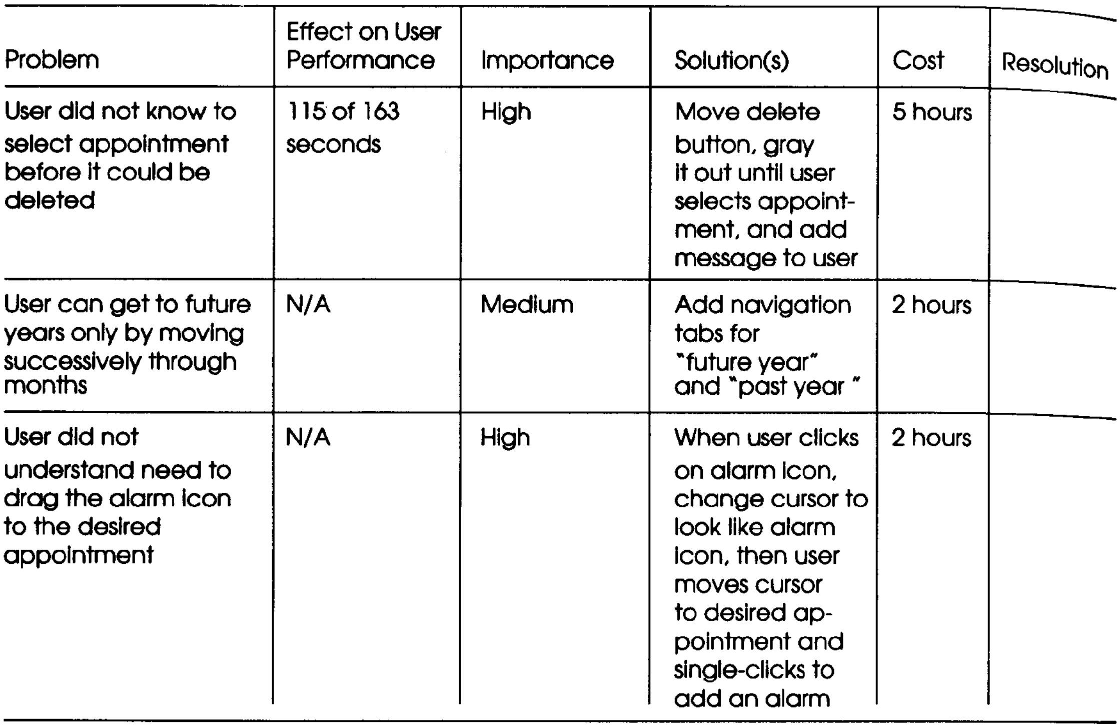
  Video/Audio tapes
- Critical incident taking
- Structured interviews

## Analyzing the Data



| Usability Attribute | Measuring Instrument | Value to be Measured | Current Level | Worst Acceptable Level | Planned Target Level | Best Possible Level | Observed Results |
|---|---|---|---|---|---|---|---|
| Initial performance | "Add appointment" task per Benchmark 1 | Length of time to successfully add appointment on first trial | 15 seconds (manually) | 30 seconds | 20 seconds | 10 seconds | P1=33 secs<br>P2=42 secs<br>P3=29 secs<br>Mean = 35 seconds |
| Initial performance | "Add appointment" task per Benchmark 1 | Number of errors on first trial | 0 errors (manually) | 3 errors | 2 errors | 0 errors | P1=2<br>P2=4<br>P3=1<br>Mean= 2.3 errors |
| Initial performance | "Delete appointment" task per Benchmark 2 | Length of time to successfully delete appointment on first trial | 12 seconds | 20 seconds | 12 seconds | 8 seconds | P1=71 secs<br>P2=42 secs<br>P3=50 secs<br>Mean= 54 seconds |
| Initial performance | "Delete appointment" task per Benchmark 2 | Number of errors on first trial | 0 errors | 4 errors | 3 errors | 0 errors | P1=5<br>P2=5<br>P3=3<br>Mean= 4.3 errors |

Resolving usability problems based on the results of your study



| Problem | Effect on User Performance | Importance | Solution(s) | Cost | Resolution |
|---|---|---|---|---|---|
| User did not know to select appointment before it could be deleted | 115 of 163 seconds | High | Move delete button, gray it out until user selects appointment, and add message to user | 5 hours | |
| User can get to future years only by moving successively through months | N/A | Medium | Add navigation tabs for "future year" and "past year " | 2 hours | |
| User did not understand need to drag the alarm icon to the desired appointment | N/A | High | When user clicks on alarm icon, change cursor to look like alarm icon, then user moves cursor to desired appointment and single-clicks to add an alarm | 2 hours | |